SUMMARY DOCUMENTATION

FOR EMAP DATA:

GUIDELINES FOR THE

INFORMATION MANAGEMENT DIRECTORY

Prepared for

Dr. Robert Shepanek
Office of Modeling, Monitoring Systems
and Quality Assurance
U.S. Environmental Protection Agency
401 M Street, S.W.
Washington, DC 20460

Prepared by

Jeffrey B. Frithsen
Donald E. Strebel
Versar, Inc.
9200 Rumsey Road
Columbia, MD 21045-1934

30 April 1995

The suggested citation for this report is:

Frithsen, J.B. and D.E. Strebel. 1995. Summary Documentation for EMAP Data: Guidelines for the Information Management Directory. 30 April 1995. Report prepared for U.S. Environmental Protection Agency, Environmental Monitoring and Assessment Program (EMAP), Washington, DC. Prepared by Versar, Inc., Columbia, MD.

EXECUTIVE SUMMARY

The Environmental Monitoring and Assessment Program (EMAP) is an interagency effort coordinated by the U.S. Environmental Protection Agency and designed to collect information to assess the condition of the nation's ecological resources. The Information Management System for EMAP was developed to capture, preserve, and provide to users data and information collected and prepared by the program. This report describes the data directory component of the EMAP Information Management System. Together with the data catalog and dictionary, the directory is one of three components that provides information about data (metadata) to users. The directory contains information that summarizes the contents of data sets and is one of the principal means by which users will identify, select, and locate EMAP data.

This document presents a summary of the metadata components for the EMAP Information Management System, design requirements for the directory, and the directory structure that was developed in response to those requirements. The requirements for the directory were determined by EMAP assessment scientists and other users through multiple joint application design sessions and feedback provided through EMAP task group information managers. In response to these requirements, the directory was designed as an integral component of the EMAP relational data base. The structure of the directory is, therefore, a set of fields organized in relational tables.

This report also contains guidelines for the compilation of directory entries. An Oracle form is presented to assist writing directory entries. The form directly interfaces to the database thus providing menus for valid field entries. The form also minimizes the need to reenter information already in the data base due to links between data base directory fields and the EMAP contact data base.

ACKNOWLEDGMENTS

The design of the EMAP data set directory presented in this document was developed with input provided by the EMAP Task Group Information Managers and other members of the EMAP Information Management team. Lawrence Cooley, Steve Hale, Melissa Hughes, Chuck Liff, Kathy Moore, and Jeff Rosen provided valuable comments on earlier drafts of this document.

Information management specialists at Technology Planning and Management Corporation (TPMC) in Durham, NC were responsible for revising and implementing the physical design of the data set directory presented in earlier drafts. Paul Cole was largely responsible for the initial implementation of the physical design and constructed the ORACLE Form to facilitate building directory entries.

The production of this report was funded by the United States Environmental Protection Agency (Environmental Monitoring and Assessment Program, Office of Research and Development) under contract No. 68-DO-0093 to Versar, Inc. It has not been subjected to the Agency's peer and administrative review and it has not been approved for publication as an EPA document.

TABLE OF CONTENTS

EXECUTIVE SUMMARY iii

ACKNOWLEDGEMENTS iv

1.0 INTRODUCTION 1

1.1 METADATA COMPONENTS 1
1.2 DIRECTORY ENTRIES 5

2.0 REQUIREMENTS DEFINED FOR THE EMAP DIRECTORY 6

2.1 DEFINITION OF USER REQUIREMENTS 6

3.0 DIRECTORY DESIGN AND GUIDELINES 9

3.1 DATA SET DEFINITION 9
3.2 DIRECTORY DESIGN 10

3.2.1 Data Set Identification Fields 11
3.2.2 Temporal Fields 13
3.2.3 Geographic Extent (Coverage) Fields 13
3.2.4 Data Center and Contact Fields 15
3.2.5 Data Set Origin and Availability Fields 21
3.2.6 Data Set Electronic Implementation 23
3.2.7 Parameter and Keyword Fields 23
3.2.8 Data Abstract Field 24

4.0 EXAMPLES OF DIRECTORY ENTRIES 26

4.1 EXAMPLE IN DIRECTORY INTERCHANGE FORMAT FORM 26
4.2 DIRECTORY SEARCH AND INFORMATION DISPLAY FORMS 28

5.0 COMPILING DIRECTORY ENTRIES 33

6.0 DIRECTORY DATA BASE DESIGN 38

7.0 REFERENCES 40

1.0 INTRODUCTION

The U.S. Environmental Protection Agency (USEPA) is coordinating the Environmental Monitoring and Assessment Program (EMAP), a multiagency effort to establish a national monitoring program designed to collect the information necessary to assess the condition of the nation's ecological resources. The purpose of this document is to provide information managers with a guide to writing data documentation (metadata) for data sets included in the EMAP information management system focusing on the structure and composition of the data set directory. The directory is one of the primary tools that will assist users with identifying and locating data sets of interest. Detailed descriptions of the metadata components included in a data set directory are provided in this document along with the requirements that led to the selection of these components and the design of the directory. These descriptions represent revisions of two earlier documents distributed to the EMAP Information Management Task Group to stimulate thoughts on metadata development (Strebel and Frithsen 1991; Thoreson et al. 1992).

The EMAP data set directory is designed to take advantage of relational data base management technology. Relational technology will facilitate the storage and maintenance of data and will minimize redundant data storage. This is particularly important since the data set directory contains information that will logically be a part of other data bases within the EMAP information management system (e.g., personnel contact information).

Maintaining metadata information within the EMAP relational data base will also provide direct links between metadata and data. These links are important because as metadata becomes more detailed (e.g., documentation concerning quality control audits), the distinction between metadata and data becomes blurred. Linkages between metadata and data are also important because users will enter the EMAP information management system both from the directory (top-down searches) and the data base (bottom-up) and will need access to both data and metadata.

1.1 METADATA COMPONENTS

Information that describes and documents data is called metadata (literally data about data). Metadata enhances the value of data by enabling assessment scientists and managers to understand the conditions, assumptions and methods under which data were collected and compiled. Metadata components also provide users with the tools necessary to identify and locate data sets of interest.

The development of metadata components for the EMAP Information Management System (IMS) is a central part of the EMAP Information Management Strategic Plan (Shepanek 1994). This plan was compiled to guide the activities of the EMAP Information Management Task Group during the period 1993-1997. A fundamental requirement outlined in the strategic plan is that metadata components must be both flexible and robust to meet the needs of

EMAP users. When fully implemented, EMAP data will be collected throughout the United States and involve the resources of multiple federal agencies. Initial users of EMAP data represent resource and coordinating groups within the program. Each of these groups has developed their own information management centers and some groups have implemented multiple data centers based upon the regional implementation of monitoring efforts. In addition to these initial data users, a diverse set of users outside of the program has arisen. These users represent various government agencies, industry, academia, and private non-government organizations.

To meet the data needs of these varied users, EMAP metadata components are structured and organized so that users can easily find the information needed to select EMAP data sets, while not becoming inundated with unnecessary information. This organization is based upon identifying components of metadata that may be logically grouped based upon the expectations of data users.

Metadata components are often referred to using terms that are not uniformly defined or applied by the information management community (Strebel and Frithsen 1991). Terms such as inventory, directory, catalog, and dictionary refer to different levels of metadata, the elements of which are typically defined by individual programs and not by generally accepted guidelines.

The three principal metadata components being developed as part of the EMAP IMS are the data set directory, catalog, and dictionary. Each component provides users with different types of information needed to identify, describe, and locate data sets. The definitions recommended for EMAP for each of these metadata components (Table 1-1) are consistent with those given in the NASA Earth Sciences Lexicon (NASA 1991), and the outline for the EMAP virtual repository presented in the EMAP Information Management Strategic Plan (Shepanek 1994; USEPA 1994).

Metadata components are designed to meet the needs of data base personnel and scientists and managers that use the data. The functionalities provided by the data set directory, catalog, and dictionary for data base personnel and other users are different (Table 1-2). In general, data base personnel use metadata to index, track, and organize data; others use metadata to identify what data are available and to obtain information used to understand the data. The functions are complementary, but in general the documentation required by data base personnel should not obfuscate other users' understanding of the data.

In a previous report (Strebel and Frithsen 1991), metadata components were outlined and organized in a manner analogous to a scientific publication. This analogy has since been further developed (Strebel and Meeson 1992; Strebel et al. 1994) to reflect the importance of metadata development to the general scientific community.

Table 1-1. Working definitions for EMAP metadata components

Metadata Component	Definition
Data set directory	Summarized data set documentation. A uniform set of descriptions of a large number of data sets, containing information suitable for making an initial determination of the existence and nature of each data set (NASA 1991).
Data set catalog	Detailed data set documentation - also referred to as the scientific documentation. A uniform set of detailed descriptions of a number of data sets and related entities, containing information suitable for making a determination of the nature of each data set and its potential usefulness for a specific application (NASA 1991).
Data dictionary	Fundamental data set documentation. The data dictionary provides a short scientific description of a parameter or variable in a data set, along with format and other basic information used in storing, searching, and displaying the data item. The dictionary contains information about the contents of each table in the relational data base.

Table 1-2. Functions of metadata components

Metadata Component	Use by Data Base Personnel	Use of Others
Directory	- Index and track data	- Identify data sets - Select data sets of interest
Catalog	- Record ancillary information about data	- Obtain descriptions of the data
Dictionary	- Organize the data - Define data formats	- Select data items of interest - Understand how to use the data

Completely describing a data set is analogous to writing a manuscript for publication in a scientific journal (Strebel et al. 1994). The metadata analogy to the manuscript is the data set catalog, which is also referred to as detailed documentation. This detailed documentation includes information concerning the originators of the data set, the general purpose for which the data were collected, sampling and laboratory methods, descriptions of the data and any manipulations or transformations of the data, related quality control and quality assurance measurements, procedures necessary for data access, and references to publications that use the data set. Details for the design of the EMAP data catalog are provided in Strebel and Frithsen (1995).

A summary of the detailed documentation is provided in the data set directory. Directory entries are analogous to the abstract of a scientific paper and contain information to assist the data user in identifying data sets that may be of interest. The directory is linked to the data set catalog so that users may quickly locate additional information concerning data sets of interest. Further, directory level information helps data management personnel index and track data sets.

A section of a data set catalog is directly linked to the data dictionary and provides technical specifications for each data item. This fundamental documentation is used by data management personnel to organize the data and by data users to select data items of interest and to determine output formats.

In addition to summarized, detailed, and fundamental documentation (directory, catalog, and dictionary metadata components, respectively), auxiliary documentation may be used to store additional information related to data sets. This auxiliary documentation can include methods manuals, photographic and electronic images of field and laboratory data sheets and quality assurance audit reports, and publications that use the data set in question. The method of access to this type of information generally depends upon user defined requirements. The EMAP Strategic Plan (Shepanek 1994) outlines a design for a virtual repository linking common metadata components (directory, catalog, and dictionary) with auxiliary documentation in related data bases, thus providing users with links between related information about EMAP and EMAP data sets.

The term inventory was used in earlier metadata discussions to refer to a subset of the information contained within the data set directory. In this context, the inventory contains information used to index and track data sets only, with little summary of the contents of the data set. This definition of inventory is redundant with what is being referred to in the directory, and is in conflict with the definition embraced by other environmental science programs (NASA 1991).

The term inventory should refer to a uniform set of descriptions of elements, or granules of a data set. (A data set granule is the smallest aggregation of data that is independently managed.) The descriptions contained within the inventory provide the information needed to select the data granule of interest. Granule descriptions typically include temporal and spatial coverage, data quality indicators, and physical storage information. The contents of the inventory, therefore, should be tailored to the set of data granules to be

described. Guidance on the contents of specialized inventories will be provided in a future report.

1.2 DIRECTORY ENTRIES

The directory serves as the primary means by which EMAP information management staff can index and track data sets and provides potential users with a first look at what data are available from the EMAP information management system. Data collected at every step of EMAP monitoring and assessment activities should be included in the directory. This includes data sets pertaining to field crew activities and coordination, sample receiving, tracking, and storage, indicator development and testing, and external data sets used in the evaluation and assessment of EMAP monitoring data.

It is unlikely that all of the data referenced in the data set directory will be included in the EMAP relational data base. Data sets created by field crews or laboratory personnel are likely to be saved as raw data files and included as part of the relational data base only after verification and validation procedures have been completed. The EMAP directory should be used as the primary tool to index and track these data sets and data files.

2.0 REQUIREMENTS DEFINED FOR THE EMAP DIRECTORY

2.1 DEFINITION OF USER REQUIREMENTS

Requirements for the EMAP directory have been developed through a variety of mechanisms. These include: review of existing information management systems for environmental data, active involvement and participation with various assessment activities conducted as part of the program, interaction with the EMAP Task Group Information Managers (TGIMs), and participation in the joint application design (JAD) sessions conducted for the EMAP Information Management Proof of Concept Project (USEPA 1993a, b). Input from these mechanisms identified a preliminary view of what EMAP users expect concerning the documentation of data. This preliminary view was refined based upon the lessons learned from compiling directory entries for data sets from the EMAP Estuaries Task Group. Additional requirements for the directory may be defined as more directory entries are completed and users begin to use the directory to identify and locate data of interest.

The contents, formats, display, and functions represented in the directory component of the EMAP Information Management System (IMS) are based upon input from current information management specialists and assessment scientists. Additional refinements should be based upon input from a user advisory committee. This committee would ensure that the IMS meets and adapt to the changing needs of its users.

The following general requirements for the EMAP directory have been defined. The requirements are not presented in order of importance.

- Directory should help users identify EMAP data sets of interest

  A primary purpose of the directory must be to assist a user's ability to identify and select data of interest. The directory must provide the user with lists of data sets available through the EMAP information management system and with summarized information about each data set to assist selecting data sets of interest.

- Directory design should assist with user defined searches

  To facilitate data set selection, the user must be able to easily perform common searches using a simple query-by-example search mechanism, as well as have the ability to create more advanced ad hoc searches using multiple fields. The directory, therefore, must contain fields that can be used during user defined searches to identify subsets of data sets.

- Directory should link to other metadata components

  The directory must be linked to the detailed (scientific) data documentation (the catalog entry) for each data set. Data users employ the detailed data

documentation as a means to further select which data sets are of interest; however, the main purpose of the scientific documentation is to provide a scientist unfamiliar with the data sufficient information to understand and use the data.

- Directory should be linked to data

  The directory should be linked to the EMAP relational data base enabling users to directly access the data described in the directory entry. Similarly, users viewing the data base should have direct access to the metadata describing that data. The linkage between data and metadata is a requirement that may be possible only in specific operating environments.

- Directory should be current

  The directory should reflect current, up-to-date information from each EMAP resource group. Sufficient resources need to be allocated for the maintenance of the EMAP directory to reflect the diversity and wealth of data being collected and managed by the program.

- Directory should accommodate all types of EMAP related data

  The directory must be flexible enough to accommodate the diverse data sets that will be managed by the EMAP information management system (field data, laboratory data, geographic data, remote sensing imagery, taxonomic data, bibliographic data, etc.). Separate directory designs should not exist for different EMAP resource groups or different types of data.

- Directory information should be easily shared

  The directory should be easily shared with other environmental data base systems within the USEPA and other federal agencies. To facilitate the exchange of directory information, it is recommended that the EMAP directory be compatible with the directory interchange format (DIF) developed for the NASA Master Directory (NASA 1991; Thoreson et al. 1992).

- Directory information concerning data access

  Initially, EMAP TGIMs stated a requirement that information concerning ordering, transferring, and accessing data sets of interest be available at the directory system level (EMAP Data Directory/Catalog/Directory Workgroup - Boulder, CO). Subsequent to stating this requirement, and due to the evolving nature of the EMAP information management system within the ORACLE relational data base management system (RDBMS), the requirement for access to this type of information has been dropped. Information concerning whom to contact with regard to a particular data set will still be maintained in the data center and contact fields of the directory.

- Directory should be easy to use

  The directory should be inherently easy to use. A consistency between screens displaying groups of directory information should exist and context sensitive help should be available at all times. The directory should be designed such that the user requires no knowledge of the underlying data structure.

The development of the data documentation components of the EMAP IMS represents an ongoing and long-term effort requiring periodic feedback from users. Features and functions provided will change and improve as the information management teams for each task group continue to evolve and additional data are added to the EMAP directory.

3.0 DIRECTORY DESIGN AND GUIDELINES

The guidelines presented below for building entries for the EMAP data set directory were developed based upon the currently known user requirements and the information management requirement to design the directory within the ORACLE RDBMS. Guidelines may continue to evolve as EMAP task groups begin to compile entries for the directory.

The resources necessary to build the directory should not be underestimated. The work involves data management personnel working closely with field and laboratory crews, quality control staff, and program design and assessment scientists to ensure that documentation accurately reflects the procedures used to create the data set. Much of this activity cannot be automated and requires input and review by information management staff familiar with the needs of scientific users and an additional review by knowledgeable scientists.

3.1 DATA SET DEFINITION

A critical first step for completing directory entries is defining the data for which documentation is being prepared. The organization of data should reflect how scientists organize and use data. Scientists continue to think of data in terms of data sets despite the capabilities for data management and reorganization provided by relational data base systems; therefore, a directory entry corresponds to a data set. A data set is arbitrarily defined, but usually represents a collection of similarly related data. Each individual EMAP task group is responsible for defining what constitutes a data set. As a guide, the definition adopted for the NASA Master Directory is offered:

> Data Set - A logically meaningful grouping or collection of similar or related data. Data having mostly similar characteristics (source or class of source, processing level and algorithms, resolution, etc.) but different independent variable ranges are normally considered part of a single data set. (NASA 1991)

Based upon this definition, a data set is a logical entity defined by scientific criteria rather than a table in the relational data base. The data set may reflect the contents of a physical file (e.g., an ASCII or SAS data set), or a part of a relational data base represented by a data view comprised of components from one or more tables. The EMAP data directory tracts both types of data sets.

Data documentation efforts will be simplified by the creation of simple data sets. Compiling directory and catalog information for a data set comprised of air temperature and wind speed is straightforward. Compiling documentation for a data representing soil type, moisture content, metal chemistry, and pesticide concentrations presents challenges for the organization of catalog metadata components (Strebel and Frithsen 1995) because these types of data are derived by different methods and quality assurance procedures. The composition of data

sets should, therefore, be determined with foresight concerning the expectations of scientists and a practical consideration of data documentation needs.

3.2 DIRECTORY DESIGN

The information in the directory presents a summary of the contents of a data set and is organized in specific fields that are part of the tables included in the EMAP relational data base. The fields can be grouped into the following types of information:

- data set identification
- temporal period
- geographic extent
- data centers and contacts
- data set origin and availability
- data set summary (abstract)

The fields represented in each of these groups are listed in Table 3-1 and described in more detail below.

Table 3-1. Summary of fields included in the EMAP data set directory

Data Set Identification Task Group Data Set Identification Number Version Entry Date Revision Date	Geographic Extent Locational Keywords Location Coordinates Included Flag Maximum Latitude Minimum Latitude Maximum Longitude Minimum Longitude
Data Set Description Abstract General Keywords	
Temporal Period Start Date End Date	Availability Availability Data Set Comments
Data Center Data Center Identification Data Center Name Address 1 Address 2 Address 3 Address 4 City State Zip Country Voice Phone FAX Phone Email Address Email Network Email Additional Information Preferred Contact Position Originating Data Center Task Group Originating Data Center Identification Originating Data Center Name	Contacts Contact Title Contact Last Name Contact First Name Contact Middle Initial Contact Role Address 1 Address 2 Address 3 Address 4 City State Zip Country Voice Phone FAX Phone Mobile Phone Pager Phone

3.2.1 Data Set Identification Fields

Task Group: Name of the EMAP task group from which the data set originates. The task group is the first line in a data set directory entry.

Recommendation: This field is mandatory. EMAP task groups are referenced using a unique two digit code. Valid codes for each task group are given in Table 3-2.

Table 3-2. Valid entries for EMAP Task Groups

01	Estuaries
02	Forests
03	Surface Waters
04	Agricultural Lands
05	Rangelands
06	Great Lakes
07	Landscape Ecology
08	Wetlands
09	Assessment
10	Design and Statistics
11	Information Management
12	Indicators
13	Integration
14	Landscape Characterization
15	Logistics and Methods
16	Stressors
17	EMAP Center

Data Set ID: Number assigned by an EMAP task group identifying a data set. The data set ID is the second line in a data set directory entry.

Recommendation: This field is mandatory. The data set identification is a positive whole number. Each task group will ensure that no data set from the same task group is assigned the same number. The Data Set ID and the Task Group will be concatenated to determine a unique identifier for each data set; therefore, different task groups having the same data set identification number will not present a problem.

Comment: Task groups having multiple data centers will need to develop a procedure to ensure that there is a unique data set identification number for each data set. A potential solution to this problem is to pre-assign data set identification numbers to each data center. For example, the EMAP Estuaries Narragansett data center may be assigned numbers between 10,000 and 19,999, and the EMAP Estuaries Gulf Breeze data center may be assigned numbers between 20,000 and 29,999. Other solutions are possible and are left to the discretion of the task group information manager.

Data Set Name: Descriptive name of data set.

Recommendation: This field is mandatory. Data set names are provided by the EMAP task group from which the data set originates. Data set names should be as descriptive as possible, reflect how real users might relate to the data, and avoid the use of acronyms and abbreviations that are specific to a particular discipline. Parameters or variables may be included in the data set name when they are important identifying characteristics of the data. Titles are limited to 200 characters.

Example: Benthic Macroinvertebrate Species composition, abundance, and biomass

Version: Version number for a data set.

Recommendation: This field is mandatory. The version number should be a positive whole number.

Comment: Any change in a data set results in the creation of a new data set or an updated version of an old data set. All changes need to be documented and that documentation included as part of the data set catalog.

Entry Date: The date an entry was incorporated into the EMAP directory.

Recommendation: This field is mandatory. The entry date is used to track when a particular entry was added into the EMAP inventory/directory. The format of the date follows the guidelines outlined below for temporal fields.

Example: 1993-08-30

Rev Date: The date a directory entry was modified or updated.

Directory information will be reviewed and revised on a periodic basis. Modifications may be necessary to reflect additional information not avail

able at the time the original directory entry was created. The date of revision will be captured in this field. Note that it is possible to modify metadata entries without creating a new version of the data set.

Recommendation: This field is mandatory. The format of the date follows the guidelines outlined below for temporal fields.

Example: 1993-10-15

3.2.2 Temporal Fields

Start Date: Earliest sampling date in this data set.

Recommendation: This field is mandatory when temporally related information is included with the data set. Dates are given in the format YYYY-MM-DD, where DD is the two-digits for the date, MM is the two digits signifying the month, and YYYY is the four-digit year. Leading zeroes are used in the entry if needed. Information concerning time can be added to the end of date information using the format YYYY-MM-DD-hh:mm:ss for date with unqualified (local) time and YYYY-MM-DD-hh:mm:ssZ for dates with Greenwich Mean Time. Hours are specified as twenty-four hour (military) time. These forms are compatible with the ISO8601 standard.

In some cases, data are tracked monthly with no reference to specific dates within the month. The date format for these data will be YYYY-MM.

Example: 1990-06-19
1990-06-19-16:00:00

End Date: Latest sampling date in this data set.

Recommendation: This field is mandatory when temporally related information is included with the data set. If data continue through the present, the end date can be omitted. Dates are presented in formats similar to those given above.

Example: 1990-09-30

3.2.3 Geographic Extent (Coverage) Fields

Loc Keyword: A descriptive word or phrase used to describe the general geographic location from which a data set originates.

Recommendation: This field is optional, however, searches of the data set directory will be facilitated when this field is completed. The vocabulary of the locational keyword field is unrestricted; however, suggested keywords include: county and state names, USEPA Regions, and names of rivers, lakes, estuaries, and national and state parks.

The locational keyword field may be repeated as many times as necessary and should reflect how real users might conduct a search for the data. The length of the field is 40 characters. The use of acronyms and abbreviations specific to a particular discipline should be avoided.

Example: Virginian Province
EPA Region I, EPA Region II, EPA Region III
Massachusetts, Rhode Island, New York, New Jersey, Delaware, Maryland, Virginia

Loc Crd Incl: A field indicating if geographic coordinates (latitude and longitude) are included in this data set.

Recommendation: This field is mandatory for all data sets containing spatial data, or data collected at specific spatial locations. Recommended values are 'Y' if geographic coordinates are included in the data set, 'N' if not. Geographic coordinates marking the outer boundaries of sample locations in the data set must be provided if the value of this field is 'Y'.

Max Lat: Northernmost latitude represented by sample points in this data set.

Recommendation: This field is mandatory if the value of the 'Loc Crd Incl' field is 'Y'. The format for geographic coordinates is determined by that used in the Master Directory. In accordance with the FIPS and ANSI standards (FIPS Pub 70-1, ANSI X3.61-1986), geographic coordinates should be indicated using alphanumeric or integer forms; however, for the EMAP directory the integer form is used. For the integer representation (+,-) the plus sign (+) or minus sign (-) must immediately precede the longitude or latitude value. Latitude is expressed in decimal degrees and is positive north of the equator. A point on the equator shall be assigned to the Northern Hemisphere.

Example: +42.4567

Min Lat: Southernmost latitude represented by sample points in this data set.

Recommendation: This field is mandatory if the value of the 'Loc Crd Incl' field is 'Y'. The format for this field is that specified for the field 'Max Lat'.

Example: + 38.4567

Max Long: Easternmost longitude represented by sample points in this data set.

Recommendation: This field is mandatory if the value of the 'Loc Crd Incl' field is 'Y'. Longitude is expressed in decimal degrees and is positive east of the Greenwich meridian. A point on the prime meridian is assigned to the Eastern Hemisphere and is, therefore, preceded by a plus sign; a point on the one hundred eightieth meridian is assigned to the Western Hemisphere and is, therefore, preceded by a negative sign.

Example: -70.4566

Min Long: Westernmost longitude represented by sample points in this data set.

Recommendation: This field is mandatory if the value of the 'Loc Crd Incl' field is 'Y'. The format for this field is that specified for the field 'Max Long'.

Example: -72.4456

Comment for locational coordinate fields: Users should carefully consider the number of significant digits given in the locational coordinate fields. Adding significant digits will specify a locational precision that is not supported by the data. Deleting significant digits will decrease locational precision.

3.2.4 Data Center and Contact Fields

Data Ctr Id: Code identifying the data center where the data are physically located.

Recommendation: This field is mandatory The data center identification number is a unique two digit code. Previously defined codes are provided in Table 3-3. Codes for data centers from which data are anticipated are also provided.

Table 3-3. Valid data center identification codes.

01	EMAP - Estuaries - Narragansett, RI
02	EMAP - Agricultural Lands - Raleigh, NC
03	EMAP - Forests - Las Vegas, NV
04	EMAP - Surface Waters - Corvallis, OR
05	EMAP - Great Lakes - Duluth, MN
06	EMAP - Central - Research Triangle Park, NC
07	EMAP - Estuaries - Gulf Breeze, FL
08	EMAP - Estuaries - Charleston, SC
09	EMAP - Landscape Characterization - Research Triangle Park, NC
10	EMAP - Rangelands - Las Vegas, NV
11	USEPA - Region I - Boston, MA
12	USEPA - Region II - Edison, NJ
13	USEPA - Region III - Philadelphia, PA
14	USEPA - Region IV - Atlanta, GA
15	USEPA - Region V - Chicago, IL
16	USEPA - Region VI -
17	USEPA - Region VII -
18	USEPA - Region VIII - Denver, CO
19	USEPA - Region IX -
20	USEPA - Region X -
99	Other - Define in Data Center Name Field

Data Ctr Name: Name for this data center.

Recommendation: This field is mandatory. Data center names indicate the Agency, program, task group, and specific data center from which the data set originated. No abbreviations should be used in data center names. The maximum length of this field is 240 characters and the field should contain no hard returns.

Example: United States Environmental Protection Agency, Environmental Monitoring and Assessment Program, Estuaries Task Group, Narragansett, RI.

Comment: The data center information fields (Data Center Identification and Data Center Name) are repeated as often as necessary when multiple data centers possess the same data set. Most data will be stored at a single data center; however, in some cases data may be stored at multiple data centers to minimize the need to move data sets across the network repeatedly. In those cases, it is useful to track additional copies of the data using the data set directory.

AddressX: Street, rural route, or post office box address of this data center.

Recommendation: This field is mandatory. Up to four lines of address, each having a length of up to 40 characters, are specified for address information. Each line of information is specified as the value for a different field: Address1, Address2, Address3, and Address4.

Example: Environmental Research Laboratory
U.S. Environmental Protection Agency
27 Tarzwell Drive

City: City of mailing address of this data center.

Recommendation: This field is mandatory. Up to 30 characters can be used to provide the city.

State: Two-letter abbreviation of the state for the mailing address of this data center.

Recommendation: This field is mandatory. The state is designated by the two character code used by the U.S. Postal Service.

Zip: Zip code of mailing address of this data center.

Recommendation: This field is mandatory. Up to 10 characters can be used to provide the zip code. The five digit (short zip code) code is required; the last four digits can be omitted if not known.

Example: 02882-1197

Country: Country of mailing address of this data center.

Recommendation: This field is optional; however, if left blank the default is be USA. Up to 40 characters can be used to describe the country.

Example: USA

Voice Phone: Voice phone number, including area code, of this data center.

Recommendation: This field is optional. Up to 18 characters can be used to provide reference to the voice phone number. If an extension also exists, this can be added at the end of the phone number as 'X1234'; however, NASA DIFs do not account for extensions. Extensions will have to be removed before forwarding DIF entries to the Master Directory.

Example: 401-792-3000

FAX Phone: FAX phone number, including area code, of this data center.

Recommendation: This field is optional. Up to 18 characters can be used to provide reference to the fax phone number.

Comment: The following group of fields specifying the electronic mail address of a data center is repeated as many times as needed to reflect multiple electronic mail addresses. In general, EMAIL addresses should be given for both internal (EPA) and external (Internet) networks to be accessible to the widest range of potential users of EMAP data.

EMAIL Address: Email address for this data center organization.

Recommendation: This field is optional. Up to 80 characters are used to specify the EMAIL address.

EMAIL Network: Network identified for this email address for this data center.

Recommended: This field is optional. Up to 20 characters are used to specify the EMAIL network. Valid names for email networks are provided in Table 3-4. Other names may be added as needed.

EM Add. Info: Any additional information that is needed to access this data center organization using the email addresses noted.

Recommendation: This field is optional and is restricted to up to 80 characters.

Examples: Data center most accessible through Internet or EPA All-in-One Mail.

Table 3-4. Valid names for email networks
Bitnet
Internet
OMNET
Telemail
USEPA All-in-One
USEPA VAX

Pref Contact: The preferred contact for this data set at a specific data center.

Recommendation: The information in this field refers to a position title within the data center (task group information manager, data librarian, technical director, etc.). The identification of specific names in this field is not recommended since the most appropriate person to contact concerning a data set may change. The information in this field is linked to the EMAP contact data base to provide a specific name for the position title given. It is recommended that for information about data, the preferred contact at each data center is the Data Librarian.

Valid names for preferred contact are the same as for the contact role specified for individuals. Valid contact names are provided in Table 3-5.

When a directory entry is created or modified, it is necessary to verify that each position title is linked to a specific person in the EMAP contact data base and that the address, telephone, and EMAIL information associated with that person is correct. This link is automatic when directory entries are made using the Oracle directory entry form (see Section 4). If directory entries are being made off-line, then name, address, telephone, and EMAIL information will need to be provided for the position titles at each data center. This information will be used to complete the data set directory entry and build the EMAP contact data base.

Information concerning personnel at a data center should be provided using the address, telephone, and EMAIL fields listed above along with the fields that follow. Fields may be repeated as needed to provide information for multiple individuals. Fields should be grouped as indicated in the examples that follow in Section 4.0.

Table 3-5. Valid names for preferred contact position at data centers and contact roles for individuals.
Director, EMAP
Technical Director
Technical Coordinator
Task Group Information Manager
Data Center Information Manager
Data Base Administrator
Data Librarian
Regional Environmental Services Division Director
Principal Investigator
Quality Assurance Officer
Chief Scientist
Project Manager

The following additional fields are used to provide information about personnel at a data center.

Cont Title: Formal title of an individual.

Recommendation: This field is optional and is restricted to up to 5 characters. Valid entries for title are given in Table 3-6.

Table 3-6. Valid entries for contact title		
Dr	Miss	Mr
Miss	Prof	

Cont Lst Name: Last name of an individual.

Recommendation: This field is optional and is restricted to up to 30 characters.

Cont Frst Name: First name of an individual.

Recommendation: This field is optional and is restricted to up to 15 characters.

Cont Mid Init: Middle initial of an individual.

Recommendation: This field is optional and is restricted to up to 1 character.

Contact Role: Role that an individual has at a particular data center.

Recommendation: This field is optional and is restricted to up to 40 characters. Valid contact roles are the same as the valid names provided for the preferred contact for a data center (Table 3-5).

Mobile Phone: Telephone number, including area code, for the mobile phone associated with an individual.

Recommendation: This field is optional. Up to 18 characters are used to provide reference to the mobile phone number.

Pager Phone: Telephone number, including area code, for the pager associated with an individual.

Recommendation: This field is optional. Up to 18 characters are used to provide reference to the pager phone number.

3.2.5 Data Set Origin and Availability Fields

Ext/Int: This field is used to identify if a data set was collected as part of the monitoring activities conducted directly for EMAP.

Recommendation: This field is mandatory. Values for this field should be 'I' if the data set was collected as part of the monitoring activities conducted directly for EMAP, and 'E' if this is a data set not originally collected by EMAP.

Origin TG: Name of the EMAP task group from which the data set originates. The task group is the first line in a data set directory entry.

Recommendation: This field is mandatory. EMAP task groups are referenced using a unique two digit code. Valid codes for each task group are given in Table 3-2. This field is marked '00' for data sets originating from other programs.

Origin Ctr ID: The data center identification code for the data center from which the data set originated.

Recommendation: This field is mandatory. The origin center and the data center ID are the same when the data center is the originator of the data set. The format of this field is the same as for 'Data Ctr Id'. Origin center codes should be a two digit code defined by the task group from which the data set originated. Valid codes are provided in Table 3-3. Additional codes can be added to reflect data obtained from other sources.

Origin Ctr Name: Origin Center Name

Recommendations: This field is mandatory if the field 'Origin Ctr ID' is "Unknown" or left blank. The field is optional otherwise.

Availability: Code reflecting data availability to specific groups of users.

Recommendations: This field is mandatory. A two-digit code is used to identify the type of user who may of access to a specific data set. In most cases, the distribution of data will be unrestricted; however, data that are not fully verified, validated, and documented should have limited distribution. Valid values for the availability code are provided in Table 3-7. The field may be repeated to reflect availability to more than one group of users.

Table 3-7. Valid entries for the availability code.

10	Available to all users
30	Available to known and registered users (EMAP Ad hoc user)
40	Limited to task group registered users (specific task group defined by the Originating Task Group field)
70	Limited to task group data base administrator
90	Central data base administrator

DS Comments: Any other comments on this data set.

Recommendations: This field is optional and should be used for any other brief information that data managers and assessment scientists may use to help identify which data sets may be useful. Extensive comments

should be included as part of the detailed (catalog level) documentation. A maximum of 240 characters can be used in this field. The field should contain no hard returns.

3.2.6 Data Set Electronic Implementation

Fields used to describe the electronic implementation of data sets (network server addresses, directories, and file names) were initially included in the EMAP directory. These fields have been removed from the directory due to concerns expressed by members of the EMAP Information Management Task Group. Where appropriate, this type of information will be given in the Data Access and Distribution section (Section 10) of the detailed documentation (Catalog) associated with each data set in the directory. The only field that remains is data set type.

Impl Type: Specifies the type of electronic implementation of the data set.

Recommendation: This field is optional. The type of electronic implementation of the data set refers to the general format of the data set (i.e., SAS, ARC-INFO, ORACLE table, ASCII, WordPerfect, Lotus 1-2-3, etc.). Up to 40 characters may be used in this field. Table 3-8 provides an initial list of valid entries for this field; additional implementation types will be added as needed.

3.2.7 Parameter and Keyword Fields

The Master Directory DIF manual (NASA 1991) defines the following fields that can be used to search for particular data sets of interest: Parameter measured, discipline keywords, and location keywords. Use of parameter and discipline keywords is not currently recommended since they are constructed using controlled vocabulary that does not reflect the diversity of EMAP data sets. Possible parameter and discipline keywords will be developed from general keywords used to document EMAP data sets. These fields may be added to the directory at a later date.

Use of general keyword fields constructed with an uncontrolled vocabulary is recommended. These keywords will facilitate data set searches and will provide a starting point for the future construction of a controlled vocabulary for parameter, discipline, and locational keywords. Table 3-9 provides a list of suggested keywords; however, the authors of data directory entries are not limited to these keywords and others may be used as needed.

Table 3-8. Valid entries for data set implementation type.
ASCII formatted, comma delimited file
ASCII formatted, tab delimited file
ASCII formatted, space delimited file
SAS (Export file format)
Lotus 1-2-3, Version 1.0 (WKS)
Lotus 1-2-3, Version 2.x (WK1)
Lotus 1-2-3, Version 3.0 (WK3)
Excel (XLS)
DBase (DBS)
DBase, Version 2 (DB2)
DBase, Version 4 (DB4)
WordPerfect for DOS, Version 5.1
WordPerfect for Windows, Version 6.0

Gen Keyword: This field provides the capability of entering general keywords for searching data sets.

Recommendations: This field is mandatory because it greatly enhances the users' ability to identify data sets of interest. The general keyword field may be repeated as many times as necessary. The length of the field should be 40 characters.

3.2.8 Data Abstract Field

The data set abstract is one of the most important components of the data set directory entry. The abstract is a concise summary of the contents of the data set and should contain brief statements of important information for the potential user. There is no limit to the length of the data abstract field.

Table 3-9. Suggested list of keywords describing the contents of data sets.

Sample Media Keywords Air Sediment Soil Water Biota	Biological Community Keywords Amphibians Benthic Invertebrates Bird Diatoms Fish Mammals Plants Reptiles Trees Wetlands Zooplankton
Habitat Keywords Agricultural Land Estuary Forest Great Lakes Lake Rangeland Stream	
Water Measurement Keywords Density Depth Dissolved Oxygen Fluorescence Light Extinction Salinity Secchi Disk Depth Temperature Total Suspended Solids	Chemical Group Keywords Alkanes Acid Volatile Sulfide Metals Polynuclear Aromatic Hydrocarbons Polychlorinated Biphenyls Pesticides Total Organic Carbon
Sediment Measurement Keywords Grain Size Toxicity	Remote Sensing Keywords Aerial photography AVHRR Thematic Mapper

4.0 EXAMPLES OF DIRECTORY ENTRIES

An example for a data set from the EMAP Estuaries resource group is provided below. The example is provided in two forms. First, in the form of a directory interchange format (DIF) as specified by the NASA Master Directory (NASA 1991). This format shows values for each of the fields defined in Section 3.1. To completely meet the requirements of the DIF, EMAP field names would be replaced by the equivalent Master Directory field names. The second form for the example (Section 4.2) represents the form a user would see to search the directory for this data set, and subsequent forms that would be used to display the information to the user.

4.1 EXAMPLE IN DIRECTORY INTERCHANGE FORMAT FORM

Version:	001
Task Group:	01
Data Set ID:	9
Data Set Name:	EMAP - Estuaries Program Level Database - 1990 Virginian Province Benthic Community Data Set
Entry Date:	1994-08-09
Rev Date:	1994-08-09
Start Date:	1990-07-19
End Date:	1990-09-30
Loc Keyword:	Virginian Province
Loc Keyword:	EPA Region I, EPA Region II, EPA Region III
Loc Crd Incl:	Y
Max Lat:	+41.38
Min Lat:	+36.47
Max Long:	-77.17
Min Long:	-70.04
Group:	Data Center
Data Ctr Cd:	01
Data Ctr Name:	United States Environmental Protection Agency, Environmental Monitoring and Assessment Program, Estuaries Task Group, Narragansett, RI
Address1:	Environmental Research Laboratory
Address2:	U.S. Environmental Protection Agency
Address3:	27 Tarzwell Drive
City:	Narragansett
State:	RI
Zip:	02882
Country:	USA
Voice Phone:	N/A
FAX Phone:	N/A

Group:	Email
EMAIL Address:	N/A
EMAIL Network:	Internet
End EMAIL	
Pref Contact:	Data Librarian
End Data Center	
Group:	Contact
Contact Role:	Data Librarian
Cont Title:	Ms.
Cont Lst Name:	Hughes
Cont Frst Name:	Melissa
Cont Mid Init:	N/A
Address1:	Environmental Research Laboratory
Address2:	U.S. Environmental Protection Agency
Address3:	27 Tarzwell Drive
City:	Narragansett
State:	RI
Zip:	02882
Country:	USA
Voice Phone:	401-792-3184
FAX Phone:	401-792-3030
Mobile Phone:	N/A
Pager Phone:	N/A
Group:	Email
EMAIL Address:	MHUGHES@NARVAX.NAR.EPA.GOV
EMAIL Network:	EPA VAX Network
End EMAIL	
EM Add. Info:	N/A
End Contact	
Ext/Int:	I
Origin TG:	01
Origin Ctr CD:	01
Origin Ctr Name:	United States Environmental Protection Agency, Environmental Monitoring and Assessment Program, Estuaries Task Group, Narragansett, RI
Availability:	40
DC Comments:	N/A
Impl Type:	SAS
Gen Keyword:	Benthic
Gen Keyword:	Macrofauna
Gen Keyword:	Estuarine

Group: Abstract
The Benthic Community Data set presents summary data from three benthic grabs collected at a station location. Benthic taxa count and abundances are summarized for all taxa and for infaunal and epifaunal taxa. Mean and total biomass are reported. Mean depth of grab penetration as well as mean percent moisture and silt-clay content are also given for each station.
End Abstract

4.2 DIRECTORY SEARCH AND INFORMATION DISPLAY FORMS

A forms interface to the EMAP relational data base provides users with the capability to search the directory for data sets of interest. Figure 4-1 shows the form used to define the search using resource group names, geographic and general keywords, and spatial coordinates. Figure 4-2 shows the same form, completed for a search of benthic data sets from the EMAP Estuaries Research Group that include sampling stations in USEPA Region III. The result of that search is shown in Figure 4-3 as a list of data sets that meet the search criteria. Directory information for one of the data sets from this list is shown in Figure 4-4.

Figure 4-1. Form used for user defined searches of the EMAP data set directory.

Figure 4-2. Completed form defining a search for EMAP Estuaries data in USEPA Region III containing benthic species data.

Figure 4-3. EMAP data sets meeting the search criteria given in Figure 4-2.

Figure 4-4. Directory information for a selected data set representing benthic data collected by the EMAP Estuaries Resource Group.

5.0 COMPILING DIRECTORY ENTRIES

The EMAP directory will be constructed and maintained in ORACLE tables. Directory entries may be made using an online form that was created and is maintained by the EMAP Information Management User Interactional Planning (UIP) Group. Alternately, directory entries can be prepared in directory interchange format (DIF) and subsequently loaded into the data base tables. The use of the online form is the preferred method of entry.

5.1 DIRECTORY ENTRY FORM

A metadata building utility using Oracle Forms 4.0 has been written to facilitate the process of building metadata entries for EMAP data sets. This form links to the data base and minimizes the need to input redundant information. Additionally, the form provides assistance with the definition of specific fields and menus containing valid entries.

Figure 5-1 shows the first screen of the directory entry form. The entry is begun by defining the EMAP task group that is responsible for the data directory entry. The menu of task groups can be viewed by clicking on the large button to the right of the field. Valid entries for other fields are accessed in a similar fashion. Users are prompted to fill-in required fields before a directory entry can be completed and committed to the data base.

Screens needed to enter additional directory information are accessed by pointing to the buttons at the bottom of the directory form. Related information in the data base can be linked to the directory entry being created through menus. A screen showing a portion of the names in the contact data base is shown in Figure 5-2. If needed, separate forms exist for the entry of contact information (Figure 5-3) and data center information (Figure 5-4). When the data base design is fully implemented, direct links will be provided between the directory and the detailed data set documentation available in the catalog.

5.2 CREATING DIF ENTRIES

Directory entries can be prepared using a word processor following the specifications outlined in the NASA Master Directory for directory interchange formats (DIF). This method may be used without direct connection to the database; however, it may lead to the production of invalid entries for specific fields. Significant editing may be needed for directory entries created using this approach.

To maintain consistency with Master Directory building efforts (NASA 1991), ASCII files should be constructed based upon the following specifications. Each line in the file begins with a field label followed by a colon (:) in the 14th position. The field value begins in the 16th position and may extend beyond 80 characters to whatever maximum is stated. The field should contain no hard returns. Fields marked "optional" may be left blank or excluded.

Figure 5-1. On-line form used to create directory entries.

Figure 5-2. Directory entry form showing menu for existing entries in the contract data base.

Figure 5-3. Form used to create or edit information for the contracts referenced through the data set directory. The same form is used to create entries for the EMAP contract data base.

Figure 5-4. Form used to create or edit information describing data centers.

6.0 DIRECTORY DATA BASE DESIGN

A physical design for the EMAP data set directory has been constructed using the ORACLE relational data base management system. The design has been physically implemented for the EMAP Information Management Proof of Concept (POC). The entity relationship diagram for this design is shown in Figure 6-1.

Figure 6-1. Entity relationship diagram for the EMAP Data Set Directory (provided by Paul Cole, TPMC).

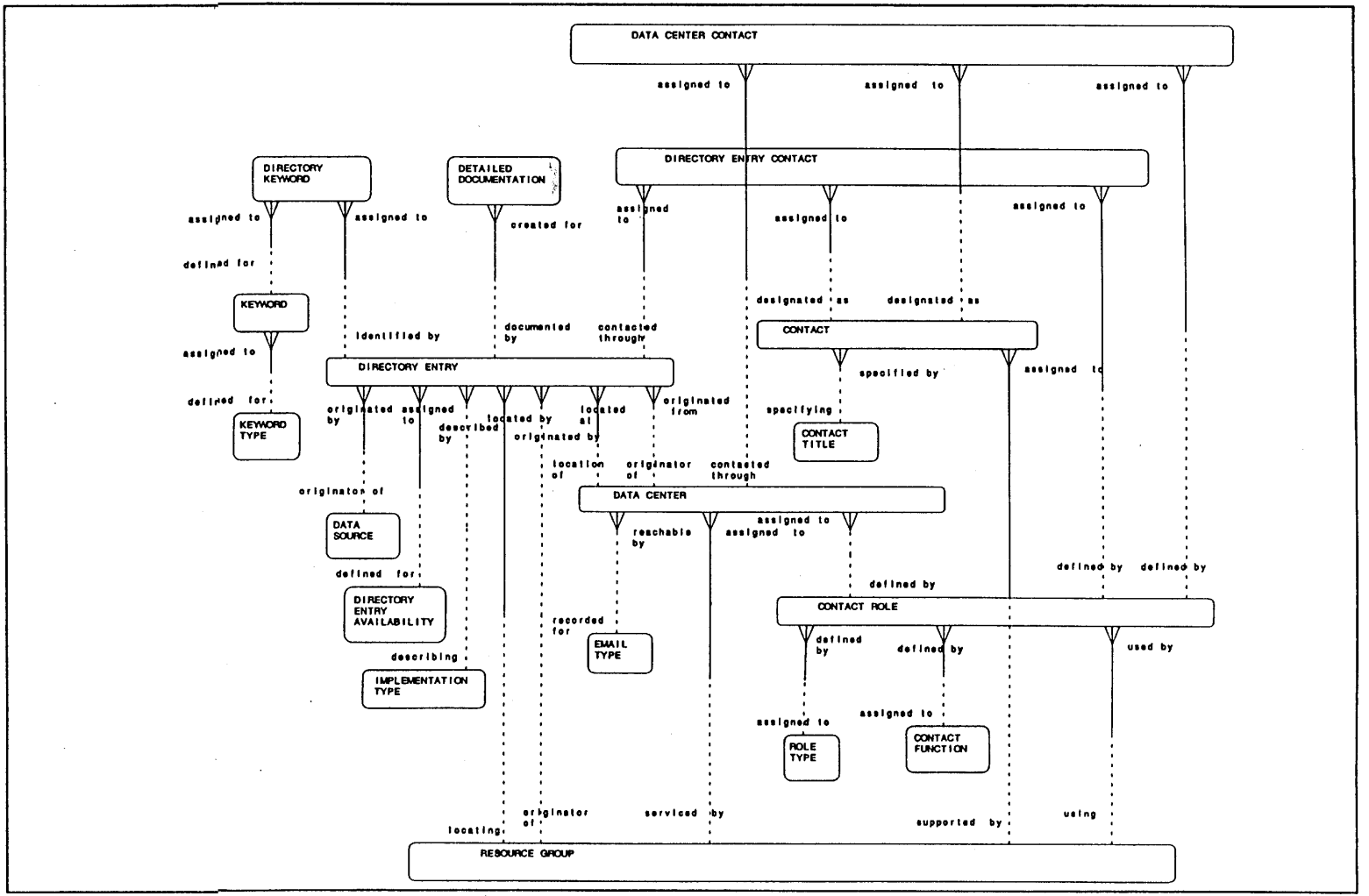

Figure 6-1. Entity relationship diagram for the EMAP Data Set Directory (provided by Paul Cole, TPMC)

7.0 REFERENCES

NASA 1991. Directory Interchange Format Manual. Version 4.0, December 1991. NASA, National Space Science Data Center, Greenbelt, MD. 9/93.

Shepanek, R. 1994. EMAP Information Management Strategic Plan: 1993-1997. EPA/620/R-94-012. U.S. Environmental Protection Agency, Office of Research and Development, Environmental Monitoring and Assessment Program (EMAP), Washington, DC.

Strebel, D.E. and J.B. Frithsen. 1991. Handling supporting information for EMAP external data sets. December 31, 1991. U.S. Environmental Protection Agency, Environmental Monitoring and Assessment Program (EMAP), Las Vegas, NV 89109. Report produced by Versar, Inc., Columbia, MD.

Strebel, D.E. and J.B. Frithsen. 1995. Guidelines for Distributing EMAP Data and Information via the Internet. April 30, 1995. Report prepared for the U.S. Environmental Protection Agency, Environmental Monitoring and Assessment Program (EMAP), Washington, DC. Report prepared by Versar, Inc., Columbia, MD.

Strebel, D.E. and J.B. Frithsen. 1995. Scientific Documentation for EMAP Data: Guidelines for the Information Management Catalog. Draft April 30, 1995. Report prepared for the U.S. Environmental Protection Agency, Environmental Monitoring and Assessment Program (EMAP), Washington, DC. Report prepared by Versar, Inc., Columbia, MD.

Strebel, D.E. and B.W. Meeson. 1992. Metadata Standards and Concepts for Interdisciplinary Scientific Data Systems. November 24, 1992. Manuscript submitted for the Proceedings for the Scientific Data Management Workshop, November 3-5, 1992, Salt Lake City, UT. Sponsored by the U.S. Department of Energy and the U.S. Environmental Protection Agency.

Strebel, D.E., B.W. Meeson, and A.K. Nelson. 1994. Scientific information systems: A conceptual framework. *In:* Environmental Information Management and Analysis: Ecosystem to Global Scales. W.K. Michener, J.W. Brunt, and S.G. Stafford, eds. Taylor and Francis, Bristol, PA.

Thoreson, H.F., D.E. Strebel and J.B. Frithsen. 1992. User requirements for data interchange formats (DIF) relevant to EMAP data set supporting information. September 30, 1992. Report produced for the U.S. Environmental Protection Agency, Environmental Monitoring and Assessment Program (EMAP), Las Vegas, NV. Report prepared by Versar, Inc., Columbia, MD.

USEPA. 1993a. Summary of the Proof of Concept Joint Application Design (JAD) Session II. January 15, 1993. U.S. Environmental Protection Agency, Office of Research and Development, Environmental Monitoring and Assessment Program (EMAP), Washington, DC.

USEPA. 1993b. System Design Specifications for the Proof of Concept (POC). February 26, 1993. U.S. Environmental Protection Agency, Office of Research and Development, Environmental Monitoring and Assessment Program (EMAP), Washington, DC.

USEPA. 1994. EMAP Information Management Virtual Repository. Draft September 9, 1994. U.S. Environmental Protection Agency, Office of Research and Development, Environmental Monitoring and Assessment Program (EMAP), Washington, DC.